Cycle3D: High-quality and Consistent Image-to-3D Generation via Generation-Reconstruction Cycle

Zhenyu Tang[1*], **Junwu Zhang** [1*], **Xinhua Cheng** [1], **Wangbo Yu** [1], **Chaoran Feng** [1]
Yatian Pang[1, 2], **Bin Lin** [1], **Li Yuan** [1 †]

[1] School of Electronic and Computer Engineering, Peking University
[2]National University of Singapore
{zhenyutang, junwuzhang, chengxinhua, wangboyu, chaoran.feng, linbin.ece}@stu.pku.edu.cn
yatian_pang@u.nus.edu, yuanli-ece@pku.edu.cn

Abstract

Recent 3D large reconstruction models typically employ a two-stage process: first generate multi-view images by a multi-view diffusion model, and then utilize a feed-forward model to reconstruct images to 3D content. However, multi-view diffusion models often produce low-quality and inconsistent images, adversely affecting the quality of the final 3D reconstruction. To address this issue, we propose a unified 3D generation framework called **Cycle3D**, which cyclically utilizes a 2D diffusion-based generation module and a feed-forward 3D reconstruction module during the multi-step diffusion process. Concretely, 2D diffusion model is applied for generating high-quality texture, and the 3D reconstruction model produces refined results with guaranteed multi-view consistency. Moreover, 2D diffusion model can further control the generated content and inject reference-view information for unseen views, thereby enhancing the diversity and texture consistency of 3D generation during the denoising process. Extensive experiments demonstrate the superior ability of our method to create 3D content with high-quality and consistency compared with state-of-the-art baselines.

Introduction

The presence of high-quality and diverse 3D assets is essential across various fields, such as robotics, gaming, and architecture. Traditionally, the creation of these assets has been a labor-intensive manual process, necessitating proficiency with complex computer graphics software. Consequently, the automatic generation of diverse and high-quality 3D content from single-view images has emerged as a crucial objective in 3D computer vision.

With the emergence of large-scale 3D datasets (Deitke et al. 2023, 2024; Yu et al. 2023; Wu et al. 2023), recent research (Xu et al. 2024a; Wei et al. 2024; Li et al. 2023; Wang et al. 2024; Xu et al. 2024b; Tang et al. 2024) has focused on large 3D reconstruction models. These models typically combine multi-view diffusion models and sparse-view reconstruction models to directly predict 3D representations (Triplane-NeRF (Shue et al. 2023; Chan et al. 2022), and 3D Gaussian Splatting (Kerbl et al. 2023)), enabling efficient 3D generation in a feed-forward manner.

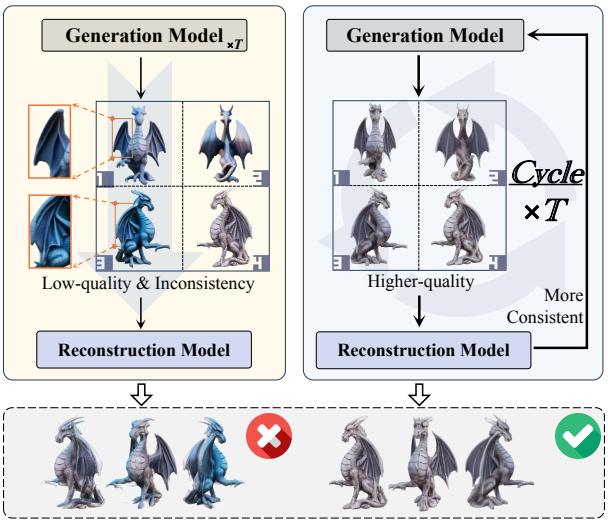

Figure 1: **Motivation of our pipeline.** Current large-scale reconstruction models often produce geometric artifacts and blurry textures due to the limited quality and consistency of the received images generated by multi-view diffusion models. Our Cycle3D cyclically uses a 2D diffusion-based generation model and reconstruction model during the multi-step diffusion process. During denoising, 2D generation model gradually improves image quality, while the reconstruction model progressively enhances 3D consistency.

However, we have observed that existing methods often encounter following two issues as shown in Figure 1: (1) **Low quality**: Multi-view diffusion models and reconstruction models are trained on limited synthetic 3D datasets, resulting in low-quality generation and poor generalization to real-world scenarios. (2) **Multi-view inconsistency**: Multi-view diffusion models struggle to generate pixel-level consistent multi-view images, while reconstruction models are typically trained on 3D consistent ground truth multi-view images. Consequently, inconsistent multi-view images usually significantly affect reconstruction results, leading to geometric artifacts and blurry textures.

*These authors contributed equally.
†Corresponding author.

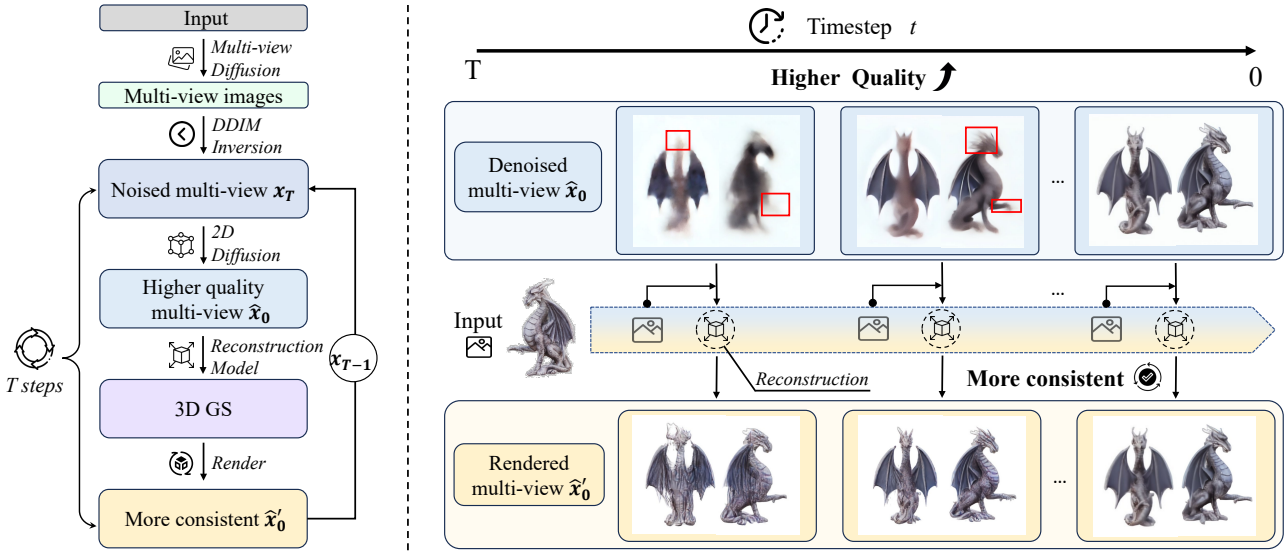

Figure 2: **Overview of our Cycle3D.** The left side illustrates the Cycle3D workflow, while the right side visualizes the denoising process. During the multi-step denoising process, the input view remains clean, the pre-trained 2D generation model gradually produces multi-view images with higher quality, while the reconstruction model continuously corrects their 3D inconsistencies. The red boxes highlight inconsistencies between the multi-view images, which are then corrected by reconstruction model.

To address these challenges, In this paper, we propose **Cycle3D**. Our method is designed based on the following two key insights: (1) The pre-trained 2D diffusion model trained on billions of web images can generate high-quality images, which is beneficial to 3D reconstruction; (2) The reconstruction model can ensure consistency across multi-views and inject consistency in 2D diffusion generation. Specifically, as shown in Figure 2, we propose a unified image-to-3D generation framework that cyclically utilizes a pre-trained 2D diffusion model and a feed-forward 3D reconstruction model during multi-step diffusion process. First, we inverse the multi-view images generated by multi-view diffusion into the initial noise, serving as shape and texture guidance. Then, in each denoising step, multi-view images are denoised and reconstructed to 3D-GS to be re-rendered, forming a loop to continue multi-step denoising. During the denoising process, the 2D diffusion model gradually provides higher quality multi-view images, while the reconstruction module progressively corrects 3D inconsistencies across multi-views. The reconstruction model can further enhance the reconstruction quality through interaction with features in the 2D Diffusion model. Additionally, 2D diffusion can control the generation of unseen views and inject reference-view information during the denoising process due to the advanced development, which further enhances the diversity and consistency of 3D generation.

We conducted extensive qualitative and quantitative experiments to validate the efficacy of our proposed Cycle3D. The experimental results demonstrate that Cycle3D outperforms other feed-forward methods and even surpasses some optimization-based methods on image-to-3D tasks. In summary, Our main contributions can be summarized as follows:

- We propose a unified image-to-3D generation framework, **Cycle3D**, which cyclically uses 2D diffusion model and a 3D reconstruction model during multi-step diffusion process. During this process, 2D diffusion model improves the quality of multi-view images, and the reconstruction model guarantees 3D consistency. The feature interaction between 2D diffusion and reconstruction model further improves the reconstruction quality.
- Leveraging the 2D diffusion model, Cycle3D can control the generation of unseen views and inject reference-view information, thereby enhancing the diversity and texture consistency of 3D generation.
- Our experiments demonstrate that our framework surpasses existing methods, achieving satisfactory image-to-3D generation with high-quality and 3D consistency.

Related Works

3D Generation from One Image

3D generation from a single image is a crucial task in computer vision, which is mainly divided into two approaches: (1) Optimization-Based Methods: These methods optimize 3D representation using 2D or multi-view diffusion models for Score Distillation Sampling (SDS) (Poole et al. 2022; Tang et al. 2023b; Qian et al. 2023; Tang et al. 2023a; Zhang et al. 2023; Yu et al. 2024; Cheng et al. 2023; Huang et al. 2023). They iteratively optimize the 3D representation of every object but are computationally intensive, leading to long optimization time. (2) Feed-Forward Generation Methods: These methods generate 3D models in a single forward pass, offering faster generation speed (Tang et al. 2024; Hong et al. 2023; Xu et al. 2024b; Wang et al. 2024; Li et al. 2023; Xu et al. 2024a; Jiang et al. 2023). These methods provide a quicker alternative to optimization-based approaches, bal-

ancing speed and quality. Our work also involves generating high-quality and consistent 3D models in the feed-forward manner, which takes only 25 seconds.

Large Reconstruction Model for 3D Generation

These methods (Tang et al. 2024; Hong et al. 2023; Xu et al. 2024b; Wang et al. 2024; Li et al. 2023; Xu et al. 2024a; Jiang et al. 2023) typically involve multi-view diffusion models (Shi et al. 2023; Wang and Shi 2023) to generate multi-view images, followed by the feed-forward reconstruction model to obtain the 3D representation. LGM (Tang et al. 2024) uses U-Net as the 3D reconstruction model, while LRM (Hong et al. 2023) and GRM (Xu et al. 2024b) employs transformers. The generative capability mainly stems from the multi-view diffusion model, with large reconstruction model primarily focusing on faithful 3D reconstruction. However, multi-view diffusion model cannot guarantee 3D consistency, leading to reconstruction artifacts. Our method uses a unified multi-step diffusion with a generation and reconstruction cycle, employing the reconstruction model to continuously correct inconsistency and 2D diffusion to progressively enhance image quality, resulting in high-quality, consistent 3D models.

Preliminary: Gaussian Splatting

Gaussian Splatting (Kerbl et al. 2023) introduces an innovative approach for synthesizing new views and fitting 3D scenes, achieving real-time rendering. Gaussian Splatting employs a set of anisotropic 3D Gaussians, to represent the scene. Specifically, each Gaussian is composed of its 3D position $p \in \mathbb{R}^3$, 3D scale $s \in \mathbb{R}^3$, color $c \in \mathbb{R}^3$, opacity $\alpha \in \mathbb{R}$, and a rotation quaternion $q \in \mathbb{R}^4$. These Gaussians can be splatted onto the image plane and rendered in real time via the differentiable tiled rasterizer.

Method

Given an RGB image, Cycle3D aims to generate high-quality and consistent 3D objects using generation model and reconstruction model. The pre-trained 2D diffusion model exhibits powerful image generation capabilities but suffers from poor multi-view image consistency. In contrast, reconstruction model reconstructs 3D objects and can render multi-view images with accurate 3D consistency. Therefore, as illustrated in Figure 3, we develop a **generation-reconstruction cycle**, cascading 2D diffusion-based **generation model** and 3D **reconstruction model** into an iterative diffusion pipeline, where 2D diffusion enhances quality and reconstruction model corrects inconsistencies. By performing generation-reconstruction cycle at each timestep during denoising process, our method achieves high-quality and consistent image-to-3D generation.

Prior-injected Generation Model

Recent image-based multi-view diffusion models (Kim et al. 2024; Zuo et al. 2024) are usually trained on limited synthetic 3D data, which hinders their ability to capture fine texture details and generalize to real-world scenarios. Therefore, we employ a 2D diffusion model (Rombach et al. 2022) (Stable Diffusion 1.5) trained on a large number of web images to generate high-quality multi-view images. Specifically, during inference, we first use the multi-view diffusion model (Kim et al. 2024) to obtain multi-view as the basic shape guidance, then inverse multi-view images to noise by performing DDIM (Song, Meng, and Ermon 2020) . The 2D diffusion model, through class-free guidance denoising, improves the quality of multi-view generation.

The 2D diffusion model effectively aligns with text prompts, so we can use more diverse text prompts to control the generation of regions not visible in the input view during the denoising process of multi-view images. Unlike directly using results generated by the image-based multi-view diffusion model, our approach allows us to achieve more diverse 3D generation by using the customized text prompt. Furthermore, benefiting from the advanced development of 2D diffusion technology, we incorporate reference-view attention features into the 2D diffusion denoising process inspired by (Cao et al. 2023) to inject reference-view prior into the generation process of multi-view images. By concatenating attention keys and values of non-input views and the reference view, we can obtain more consistent textures in multi-view images, improving the quality of 3D generation.

In our framework, the 2D diffusion model does not independently complete the entire denoising process. Within the generation-reconstruction loop, we directly estimate $\hat{\mathbf{x}}_0$ from the noise predicted by the 2D Diffusion model at every timestep, which is then used for the following 3D reconstruction. We represent this process as follows:

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, y, t)). \tag{1}$$

where $\boldsymbol{\epsilon}_\theta$ denotes 2D diffusion model, $\bar{\alpha}_t$ schedules the amount of noise added at timestep t, with y representing the text description. Subsequently, we use the frozen VAE to transform $\hat{\mathbf{x}}_0$ from the latent space to the image space.

Time-sequential Reconstruction Model

In the Cycle3D framework, we utilize a feed-forward 3D reconstruction model to predict attributes of 3D Gaussians from the multi-view $\hat{\mathbf{x}}_0$ obtained via 2D diffusion model at the each timestep of the unified diffusion process. Here, we employ an asymmetric U-Net Transformer $\mathcal{G}_\phi$ as proposed in (Tang et al. 2024), which predicts pixel-aligned Gaussian parameters from the feature of each pixel in the final layer of the U-Net. Benefiting from the differentiable real-time rendering of Gaussian Splatting, reconstruction model can be integrated into our diffusion framework for end-to-end training, enabling efficient fine-tuning.

In the denoising process of the 2D diffusion model, different timesteps produce varying levels of noise, which in turn affect the image qaulity of the directly estimated $\hat{\mathbf{x}}_0$. Therefore, to enhance the robustness of the model in reconstructing $\hat{\mathbf{x}}_0$ at different timestep, we insert zero-initialized projection layers into each ResNetBlock within the U-Net. These layers map the time embeddings from the 2D Diffusion model to fit the reconstruction model. This creative adjustment helps the reconstruction model adapt to the $\hat{\mathbf{x}}_0$ esti-

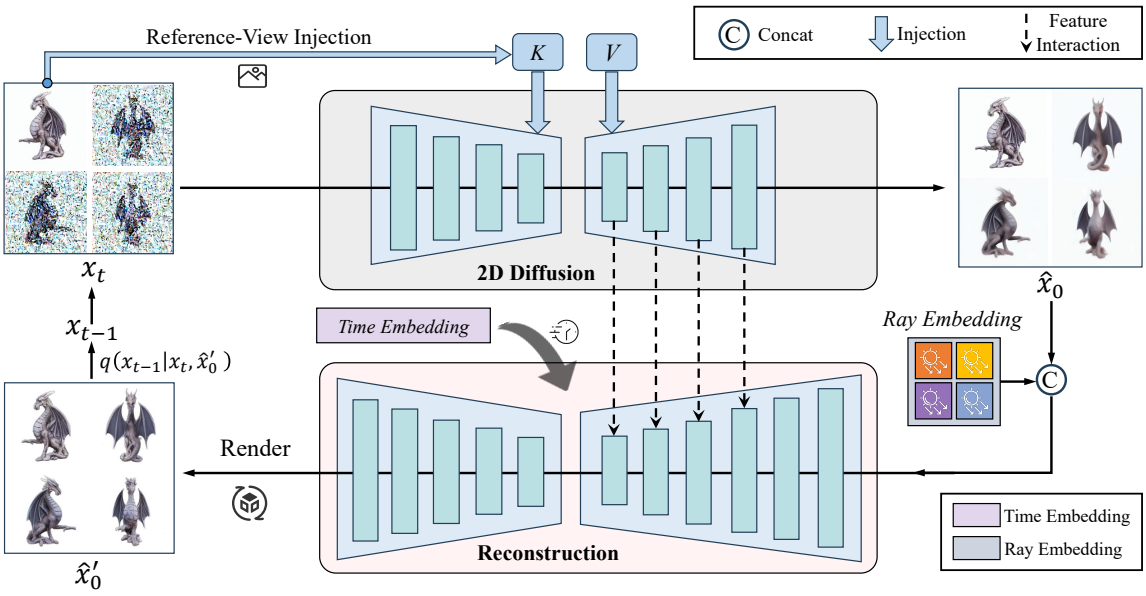

Figure 3: **Process of our Cycle3D.** We propose a unified image-to-3D Diffusion framework that cyclically utilizes pre-trained 2D Diffusion model and 3D reconstruction model. During denoising, 2D Diffusion model can inject reference-view priors, and the reconstruction model incorporates time embeddings to adapt to $\hat{\mathbf{x}}_0$ at different timesteps. Additionally, the interaction between features of reconstruction model's encoder and 2D Diffusion model's decoder enhances robustness of reconstruction. During inference, we use the multi-view images $\hat{\mathbf{x}}_0'$ rendered by reconstruction model and the previous step $\mathbf{x}_t$, resampling to obtain $\mathbf{x}_{t-1}$ to continue generation-reconstruction cycle, while keeping the reference view clean.

mated at different timestep, thereby significantly enhancing the quality of time-sequential 3D reconstruction.

To further tune the reconstruction model to adapt to our enhancements, we supervise the training using T images $\hat{\boldsymbol{I}}$ and alpha masks $\hat{\boldsymbol{M}}$ rendered by $\mathcal{G}_\phi$ with the corresponding ground truth $\boldsymbol{I}$ and $\boldsymbol{M}$. The loss function is as followers:

$$\mathcal{L}_{total} = \sum_{t=1}^{T} \left(\mathcal{L}_{\text{img}}(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_t) + ||\hat{\boldsymbol{M}}_t - \boldsymbol{M}_t||_2 \right), \tag{2}$$

$$\mathcal{L}_{\text{img}}(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_t) = ||\hat{\boldsymbol{I}}_t - \boldsymbol{I}_t||_2 + \lambda * \mathcal{L}_{\text{LPIPS}}(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_t), \tag{3}$$

where $\mathcal{L}_{\text{LPIPS}}$ is a perceptual image patch similarity loss (Zhang et al. 2018), and the weight λ is set to 0.50.

Generation-Reconstruction Cycle

To cascade 2D diffusion model and reconstruction model into the generation-reconstruction cycle, instead of inputting the denoised latent $\mathbf{x}_{t-1}$ into reconstruction model, we decode predicted $\hat{\mathbf{x}}_0$ using the VAE decoder to obtain clean multi-view images for reconstruction model. This aligns with the pretraining inputs of the reconstruction model, reducing the gap and easing the joint fine-tuning. Additionally, instead of using 2D diffusion model output $\hat{\mathbf{x}}_0$ to perform the DDIM sample step $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_0)$ for updating $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$, we adopt multi-view images $\hat{\mathbf{x}}_0'$ rasterized by 3D Gaussians output of the reconstruction model at the timestep t from the same observing views as $\hat{\mathbf{x}}_0$:

$$\mathbf{x}_{t-1} = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0' + \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t, \tag{4}$$

where $\beta_t = 1 - \alpha_t$. Compared to $\hat{\mathbf{x}}_0$, $\hat{\mathbf{x}}_0'$ has accurate 3D consistency, making the sampling trajectory more 3D consistent. Consequently, the final denoised result $\mathbf{x}_0$ is more consistent and of higher quality than the result denoised only by the existing multi-view diffusion model, leading to the reconstruction of higher-quality 3D Gaussian structures.

Furthermore, training reconstruction model on a limited synthetic 3D dataset may affect its performance with real-world images. Based on the observation that reconstruction model's reconstruction process can be understood as a sequence from multi-view images to multi-view features, and finally to multi-view Gaussians, we can enhance the reconstruction process using features from the 2D diffusion model pretrained on a large number of web images. Specifically, we introduce zero-initialized cross-attention layers in the encoder block of reconstruction model to interact the decoder features of the 2D diffusion model with the encoder features of the reconstruction model, forming a U-Net structure. This innovative modification makes the reconstruction model more robust when reconstructing real-world images.

Experiment

Implementation Details

Datasets. We use the G-objaverse dataset (Qiu et al. 2024) to train our model. Derived from the original Objaverse (Deitke et al. 2023), G-objaverse excludes 3D models with poor captions and includes a large number of high-quality renderings generated through a hybrid technique involving rasterization and path tracing. We utilize a further filtered subset contain-

Methods/Metrics	PSNR↑	SSIM↑	LPIPS↓	CLIP-Similarity↑	Contextual-Dis↓
DreamGaussian (Tang et al. 2023a)	19.4900	0.8311	0.1145	0.7136	1.8139
Wonder3D (Long et al. 2024)	18.0926	0.8164	0.1764	0.7596	1.7914
One-2-3-45 (Liu et al. 2024)	14.0064	0.7405	0.3976	0.6363	2.1069
TriplaneGaussian (Zou et al. 2024)	18.4044	0.8284	0.1515	0.7399	1.7803
OpenLRM (Hong et al. 2023)	18.6433	0.8301	0.1255	0.7567	1.7037
LGM (Tang et al. 2024)	18.6909	0.8360	0.1417	0.7990	1.6504
Cycle3D (Ours)	**20.2452**	**0.8729**	**0.1117**	**0.8238**	**1.6031**

Table 1: We show quantitative results of image-to-3D in terms of PSNR↑ / SSIM↑, LPIPS↓ / CLIP-Similarity↑ / Contextual-Distance↓ for our test dataset. The **bold** reflects the best result for optimization-based methods and feed-forward methods.

ing approximately 80K 3D objects. Each model is rendered with 36 views, from which we randomly sample 4 views with elevation angles in the range [-5°, 5°] as input multi-views, using the first frame as the condition image. Additionally, we sample 8 views from the 36 views for extra supervision of the training process.

We combine the Realfusion15 dataset (Melas-Kyriazi et al. 2023) with the dataset collected by Make-It-3D (Tang et al. 2023b), using these images from diverse styles as our test dataset. Additionally, we further evaluate the 3D generation quality on 50 objects from the GSO dataset (Downs et al. 2022) with multi-view ground truth.

Experimental Settings. Our Cycle3D is trained on 8 NVIDIA A100(80G) with batch size 8 for about 1 day. We utilized the AdamW optimizer with a learning rate of 1e-4 and a weight decay of 0.05 for 30 epochs. Additionally, we followed (Tang et al. 2024) to clip the gradient with a maximum norm of 1.0 and employed BF16 mixed precision with Deepspeed Zero2 (Rasley et al. 2020)for efficient tuning. During inference, we use the DDIM scheduler (Song, Meng, and Ermon 2020), setting the sampling steps to 30, and take about 25 seconds to generate a 3D object.

Evaluation Metrics. We use PSNR, SSIM, and LPIPS (Zhang et al. 2018) to measure reconstruction quality, and CLIP score (Radford et al. 2021) and contextual distance (Mechrez, Talmi, and Zelnik-Manor 2018) to assess image similarity. The quality of 3D generation is evaluated by comparing 180 rendered views with the ground truth.

Baselines. We select baselines for comparison, including optimization-based existing image-to-3D methods: DreamGaussian (Tang et al. 2023a), Wonder3D (Long et al. 2024), and some existing feed-forward methods: One-2-3-45 (Liu et al. 2024), TriplaneGaussian (Zou et al. 2024), LRM (Hong et al. 2023), LGM (Tang et al. 2024).

Comparison

Qualitative Comparisons. We compared our approach with recent optimization-based and feed-forward based methods. For more fair comparison, Cycle3D and LGM take the same generated multi-view inputs. As shown in Figure 4, we used a wide range of wild images to evaluate the quality of image-to-3D generation, and our Cycle3D achieved the best visual results. TriplaneGaussian (Zou et al. 2024) and OpenLRM (Hong et al. 2023) fail to complete unseen regions with high quality. DreamGaussian (Tang et al. 2023a) often produces unrealistic geometry, while Wonder3D (Long et al. 2024) tends to generate blurry textures. LGM (Tang et al. 2024) often generates blurry textures and geometric artifacts like floating 3D Gaussian splats, due to low-quality and inconsistent multi-view images. In contrast, our method can generate high-quality and consistent 3D objects due to the proposed generation-reconstruction circle.

(I)	(II)	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	Contextual↓
✗	✗	19.2491	0.8497	0.1361	0.7986	1.6399
✓	✗	20.0198	0.8702	0.1187	0.8045	1.6348
✓	✓	**20.2452**	**0.8729**	**0.1117**	**0.8238**	**1.6031**

Table 2: Quantitative results of ablation study on our test dataset. (I) denotes the feature interaction between the 2D diffusion model and the 3D reconstruction model, and (II) represents the injection of reference-view features during the 2D diffusion denoising process.

Quantitative Comparisons. As presented in Table 1, we quantitatively evaluate the quality of the generated 3D objects for our test dataset. Notably, Cycle3D surpasses all baseline methods on all metrics, even outperforming existing optimization-based methods.

Ablation and Diverse Generation

In this section, we provide detailed quantitative and qualitative analysis, as shown in Figure 5 and Table 2. We also experimented with leveraging the text descriptions of the 2D diffusion model to control the generation of unseen regions from non-input viewpoints, as illustrated in Figure 6. The ablation study on inserting time embedding into the reconstruction model is also presented in the *Appendix*.

Effectiveness of Feature Interaction. The reconstruction model, trained only on a limited synthetic 3D dataset, often lacks the capability to accurately reconstruct complex and detailed textures in real-world scenarios, resulting in blurry textures, as depicted by the red boxes on the right side of Figure 5. The 2D diffusion model, trained on a large number of real web images, exhibits robust performance on real-world textures. The feature interaction between the encoder of the 2D diffusion model and the decoder of the reconstruction model significantly enhances the reconstruction of complex texture details, as evidenced by comparing the red-boxed areas in the last two columns of Figure 5. Addition-

Figure 4: Qualitative comparisons on image-to-3D generation. Zoom in for more details.

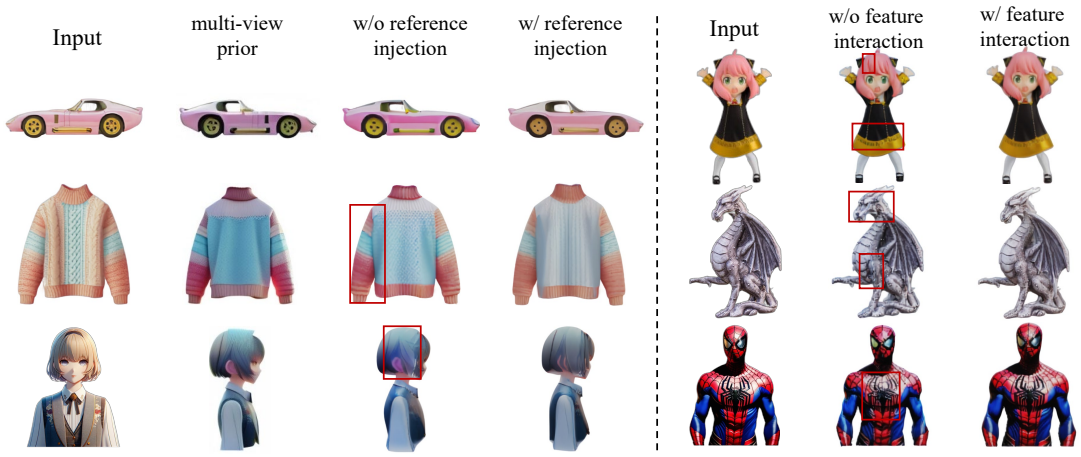

Figure 5: Qualitative ablation study by removing reference-view injection or feature interaction between 2D diffusion and reconstruction model. Multi-view prior refers to the multi-view images generated by the multi-view diffusion, used as the initial noise of 2D diffusion model through DDIM inversion. The red boxes highlight some abnormal or blurry textures. Reference-view injection can avoid abnormal textures shown in the multi-view prior that are inconsistent with the input image, while the absence of feature interaction significantly degrades the reconstruction quality.

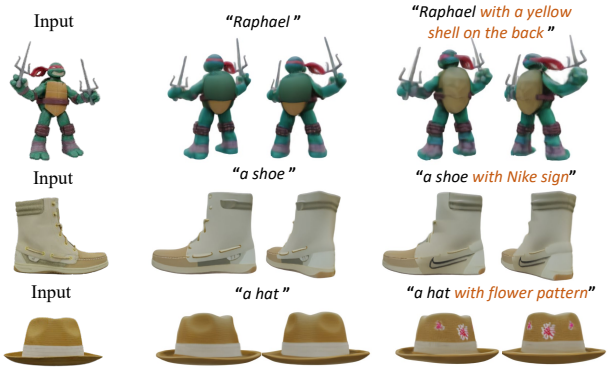

Figure 6: Diverse 3D generation by customized texts.

ally, Table 2 also demonstrates that feature interaction significantly enhances the quality of 3D reconstruction.

Effectiveness of Reference-view Injection. When multi-view diffusion generates multi-view images with unrealistic textures or textures that do not match the reference view, the 2D diffusion model, using the initial noise obtained from the multi-view images through DDIM inversion, can still produce abnormal textures inconsistent with the reference view. As shown in the third column on the left side of Figure 5, although multi-view interaction through the reconstruction model alleviates texture inconsistency to some extent, the car's body, the coat, and the girl's hair and ear still exhibit abnormal textures due to the inconsistent multi-view prior generated by the multi-view diffusion, leading to discrepancies with the reference view. By injecting information from the reference view into the 2D diffusion denoising process, we can generate multi-view textures that are more consistent with the reference view, as shown in the fourth column in Figure 5. Table 2 also proves reference-view injection can enhance the consistency of textures, as evidenced by increased CLIP similarity and reduced contextual distance.

Diverse Generation. Benefiting from the 2D diffusion model's excellent text alignment capability, we can apply diverse and customized text to control one or more non-input views, generating more varied textures in areas not visible from the reference view. As shown in Figure 6, the texture in the second column are primarily based on the multi-view prior generated by the multi-view diffusion and the injection of reference view information during the denoising process. By incorporating fine-grained textual information as conditions, we can achieve diversified and customized 3D generation, as illustrated in the last column.

Limitations

Due to the lack of large-scale 3D scene datasets, our current method is limited to object-level 3D generation and cannot be extended to scene-level generation. when large-scale scene datasets become available in the community, future work can explore more complex 3D scene generation.

Conclusion

In this paper, we introduce Cycle3D, an image-to-3D generation framework that cyclically utilizes 2D diffusion-based generation model and the 3D reconstruction model during the multi-step diffusion process. As the denoising proceeds, the 2D diffusion model progressively generates multi-view images with higher quality, while the reconstruction model gradually corrects 3D inconsistencies. 2D diffusion model can also control the generation of unseen views and inject reference-view information during denoising. Reconstruction model further interacts with 2D diffusion, enhancing the reconstruction capability. Extensive experiments demonstrate that our method surpasses existing state-of-the-art baselines in generation quality and consistency.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (No. 62332002, 62202014, 62425101).

References

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16123–16133.

Cheng, X.; Yang, T.; Wang, J.; Li, Y.; Zhang, L.; Zhang, J.; and Yuan, L. 2023. Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. *arXiv preprint arXiv:2310.11784*.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; et al. 2024. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Downs, L.; Francis, A.; Koenig, N.; Kinman, B.; Hickman, R.; Reymann, K.; McHugh, T. B.; and Vanhoucke, V. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, 2553–2560. IEEE.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Huang, Y.; Wang, J.; Shi, Y.; Tang, B.; Qi, X.; and Zhang, L. 2023. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In *The Twelfth International Conference on Learning Representations*.

Jiang, H.; Jiang, Z.; Zhao, Y.; and Huang, Q. 2023. LEAP: Liberate Sparse-view 3D Modeling from Camera Poses. arXiv:2310.01410.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4): 1–14.

Kim, S.; Shi, Y.; Li, K.; Cho, M.; and Wang, P. 2024. Multi-view Image Prompted Multi-view Diffusion for Improved 3D Generation. *arXiv preprint arXiv:2404.17419*.

Li, J.; Tan, H.; Zhang, K.; Xu, Z.; Luan, F.; Xu, Y.; Hong, Y.; Sunkavalli, K.; Shakhnarovich, G.; and Bi, S. 2023. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2024. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9970–9980.

Mechrez, R.; Talmi, I.; and Zelnik-Manor, L. 2018. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, 768–783.

Melas-Kyriazi, L.; Laina, I.; Rupprecht, C.; and Vedaldi, A. 2023. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8446–8455.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qian, G.; Mai, J.; Hamdi, A.; Ren, J.; Siarohin, A.; Li, B.; Lee, H.-Y.; Skorokhodov, I.; Wonka, P.; Tulyakov, S.; et al. 2023. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*.

Qiu, L.; Chen, G.; Gu, X.; Zuo, Q.; Xu, M.; Wu, Y.; Yuan, W.; Dong, Z.; Bo, L.; and Han, X. 2024. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9914–9925.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rasley, J.; Rajbhandari, S.; Ruwase, O.; and He, Y. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3505–3506.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Shue, J. R.; Chan, E. R.; Po, R.; Ankner, Z.; Wu, J.; and Wetzstein, G. 2023. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20875–20886.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2024. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. *arXiv preprint arXiv:2402.05054*.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2023a. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*.

Tang, J.; Wang, T.; Zhang, B.; Zhang, T.; Yi, R.; Ma, L.; and Chen, D. 2023b. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22819–22829.

Wang, P.; and Shi, Y. 2023. ImageDream: Image-Prompt Multi-view Diffusion for 3D Generation. arXiv:2312.02201.

Wang, Z.; Wang, Y.; Chen, Y.; Xiang, C.; Chen, S.; Yu, D.; Li, C.; Su, H.; and Zhu, J. 2024. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*.

Wei, X.; Zhang, K.; Bi, S.; Tan, H.; Luan, F.; Deschaintre, V.; Sunkavalli, K.; Su, H.; and Xu, Z. 2024. MeshLRM: Large Reconstruction Model for High-Quality Mesh. *arXiv preprint arXiv:2404.12385*.

Wu, T.; Zhang, J.; Fu, X.; Wang, Y.; Ren, J.; Pan, L.; Wu, W.; Yang, L.; Wang, J.; Qian, C.; et al. 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 803–814.

Xu, J.; Cheng, W.; Gao, Y.; Wang, X.; Gao, S.; and Shan, Y. 2024a. InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models. *arXiv preprint arXiv:2404.07191*.

Xu, Y.; Shi, Z.; Yifan, W.; Chen, H.; Yang, C.; Peng, S.; Shen, Y.; and Wetzstein, G. 2024b. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*.

Yu, W.; Yuan, L.; Cao, Y.-P.; Gao, X.; Li, X.; Hu, W.; Quan, L.; Shan, Y.; and Tian, Y. 2024. HiFi-123: Towards High-fidelity One Image to 3D Content Generation. arXiv:2310.06744.

Yu, X.; Xu, M.; Zhang, Y.; Liu, H.; Ye, C.; Wu, Y.; Yan, Z.; Zhu, C.; Xiong, Z.; Liang, T.; et al. 2023. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9150–9161.

Zhang, J.; Tang, Z.; Pang, Y.; Cheng, X.; Jin, P.; Wei, Y.; Ning, M.; and Yuan, L. 2023. Repaint123: Fast and High-quality One Image to 3D Generation with Progressive Controllable 2D Repainting. arXiv:2312.13271.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zou, Z.-X.; Yu, Z.; Guo, Y.-C.; Li, Y.; Liang, D.; Cao, Y.-P.; and Zhang, S.-H. 2024. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10324–10335.

Zuo, Q.; Gu, X.; Qiu, L.; Dong, Y.; Zhao, Z.; Yuan, W.; Peng, R.; Zhu, S.; Dong, Z.; Bo, L.; et al. 2024. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*.